

TF-IDF 法に基づくタグの自動付与

河上 哲也 千種康民[†] 服部泰造[‡]

東京工科大学 コンピュータサイエンス学部[†] 東京国際大学[‡]

1. 研究の背景と目的

現在、世の中には多種多様なドキュメントがあるが、それらを分類する方法はひとつの課題である。最近では、サイト運営者が分類（カテゴリ分け）するのではなく、利用者がコンテンツにタグをつけて分類する方法が登場した。

フォークソノミは、コンテンツに対して人がタグ（キーワード）を多数付加していくことで利用者本位な分類をしていく手法である。通常フォークソノミは人の手により行われるため、大変な労力がかかる。

そこで、本研究ではフォークソノミにおけるタグ付けを自動化する研究を行った。

△初音ミク ニコニコ動画まとめ ニコニコ

ニコニコのタグ検索の「精度」の話 - Myrmecoleon in Paradoxical L
館 d:id myrmecoleon
△ウェブ □API RSS [これはひどい] はてブ タグクラウド □ 2007年1
△Search niconico tag research ニコニコ動画 考察 タグ

適度にdisされるのは乾布摩擦のようなもの - シロクマの肩籠(汎適)
d:id p_shirokuma

△食 □DIS シロクマダメージ 嫌がらせ 軽蔑 □ 20 users 2007年10月24日
△プロゲ □論 web 芳川龍之介 メタ blog 心理 考察 私信

図 1 フォークソノミの例

2. 特徴

フォークソノミによる分類は、カテゴリ分けにとらわれることのない自由な分類が可能となる。タグの綴りの間違えや、同義語といった問題点を解決することができる。意味のないタグ付がなくなる。タグ付が自動化されるため、労力が軽減される。タグ付がされていない既存のテキストに適応できる。といった特徴がある。

Automatic Ttagging Based on TF-IDF
Tetsuya KAWAKAMI[†], Yasutami CHIGUSA[†] and
Taizoh HATTORI[‡]
[†]Tokyo University of Technorogy
[‡]Tokyo International University

3. 実装方法

対象ドキュメントからキーワードを抽出し、それを評価してタグを出力するシステムを実装する。

キーワードを抽出するためにドキュメントを形態素解析を用いて形態素に分割し、それらの単語を TF-IDF 法により重みづけをして評価を行う。最終的にスコアの高い単語をタグとして出力する

TF-IDF 法は単語の出現頻度 (tf) と単語の希少度 (idf) で単語を評価する手法である。スコア (w) の高い単語をそのドキュメントのタグとして出力する。

$$tfidf(d, t) = tf(d, t) \cdot idf(t) \quad (1)$$

$$idf(t) = \log\left(\frac{N}{df(t)}\right) \quad (2)$$

d:入力文書 t:対象となる単語 tf:単語の出現頻度 df:単語の文書出現頻度 N:総文書数

評価したタグはデータベースにしておき、同じタグが付けられたときや、タグによる検索が行われた時にタグを再評価する。これにより使用頻度の高いタグは評価が高くなり、使用頻度の低いタグは評価が下がる。

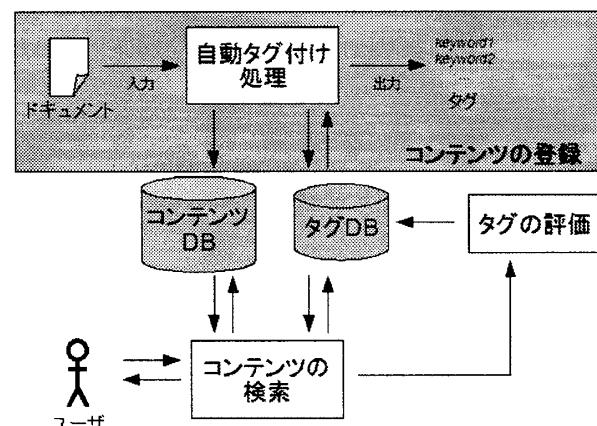


図 2 システム構成図

4. タグのスコア付けの結果

図 3-6 は §3 の手法により求めたスコアをもとに、入力文書のテキストの各単語のフォントサイズ変化をさせることで、もとの文書を視覚的にわかりやすくした例である。また、結果は品詞ごとに色を変えてある。

5. まとめ

ドキュメントの分類手法の一つであるフォークソノミをとりあげ、その問題点及びその解決方法として自動タグ付による方法を提案し、その手法についてとりあげた。また、スコアづけの評価を行うための視覚化についてとりあげた。

森の近くの草原に、うさぎたちがいた。たくさんの中、雪うさぎとほらうさぎ、黒うさぎ、色が、黒うさぎは、結構に入り込まない。だって真っ黒じゃすぐ見分けてしまうよ。危ないもの」 そう言しながら、雪うさぎたちは、黒うさぎを追いかける。「どうして森は真っ黒なんだろ? 楽しかったからかな? 楽が無いのかな?」「だから独り立ったのか?」 はもう覚えてない。こんなにも同じカタチのうさぎたちは、近づいた時に遠ざかる。だから、黒うさぎは嘘をつけた。独り立ちだって、「だいじょうぶ、だいじょうぶだいじょうぶ」 雪の日も冬の日も雪の日も、そのまま振り返して過ぎていた。そんなある冬のこと、黒うさぎは嘘の絆でちょっと泣いた。気付けば身体中苦だらけ、まるで雪うさぎみたいに真っ白だった。「大丈夫かい?」 なれど、声に振り向かず、きれいな毛並みの雪うさぎが泣いていた。確かに声をかけてやるなんて、どれだけ慈愛のことなのか。あれもこんな優しい言葉。
涙を必死にこらえて、黒うさぎは泣いた。「大丈夫だよ。ありがとう」

図 3 童話を視覚化した例

頭が痛い。ガンガンする。目を開けると仰向けに倒れていた。頭には、青緑色の液がこぼてくる。痛い体を起こしてみると、砂利道の面積はますます広がっている事が分かった。なんとか腰張って起き上がり、迷ひながら見渡すが、全く見えない。砂利道と田んぼ、真っ青な空と入道雲、そして青々とした山以外に何もない景色。「ここはどこだ…。そしてわたしはここで何をしていたのだろう…」 一体、ここはどこなのか、そして何をしていたのか、頭を悩ませてしかどうか分からぬが、まるで思ひ出せない。しかしこのままここにじっとしているのも不楽だ。とにかくこの砂利道を這いつたい、そろそろ立ち砂利道を山の方へ進んで行った。すると大きな岩で倒れている村夫らしき猫がいた。迷ひ腰をして三つ。『頭から血が止まってるよ。腹も膨れてるし、一体どうしたのかね』 いや、よく分からぬのです。一体どこでしようか? 「そんな事よりも、まずは手当てをした方がよさそうだ。家へ戻る男、ありがたく好意を受ける事になった。すぐ機に家があった。この辺には既

図 4 ショートストーリーを視覚化した例

中高一貫校相次ぎ高校募集中止「6年で進学実績 2008年01月15日 15時09分 高校からの募集をめぐる私立の中高一貫校が大都市圏で相次いでいる少子化により学校間の競争が激しくなる中、卒業後につい込み、6年一貫教育で進学実績を重視したいという意思や、中高一貫組と高校組とで授業の履歴を「二度手間」を避ける目的もある。公立中からは「優秀な子が私立に進むしまう」との現象もある。■ 大妻中野(東京都中野区)は2002年の入試から高校募集を停止する。かつては、ただだったが、少子化が進むと、高校募集しないといけないと決まりから中学募集を始めた。中学からの割合を増加してきた。高麗中学校長は「公立にも一貫校が生まれ、一貫教育の本質が改められてきている。首都圏での中学受験ブームも底堅かった」と話す。高校募集がない学校は完全中高一貫校とも呼ばれる。首都圏に約300校ある中高一貫校のうち70校ほどある。日能研グループのNTS教育研究所によると、この数年、中高連携が完全中高一貫化、中学から優秀な生徒を採用することで、公立高校の「滑り止め」から脱皮する傾向があるようだ。この傾向は募集定員にも表れている。03年度以降のデータが完全

図 5 ニュースを視覚化した例

現在、世の中には多種多様なタグが存在するが、それらを分類する方法がひとつある。最近では、サードパーティがタグを分類する方法が登場した。フォークソノミはコンテンツに対して、タグを付けて分類する方法が登場した。フォークソノミはコンテンツに対して、タグ(タグ)を多数付加することで利用者本位の分類をしていく手法である。タグソノミは人の手によってされるため、大きな労力が必要となる。そこで、本研究では自動タグソノミによる自動化を試みた。フォークソノミによる分類とは、タグを付けて分類されることで、自由なタグが可能となる。タグの繋りの言葉や、同義点を解消することができる。意味のタグ付が可能で、タグ付が自動労力が削減される。タグ付がされている既存のテキストに適応できるとする。

図 6 本稿の 2 節を視覚化した例

参考文献

- [1] 西村 悟, 千野 晋平, 三木 光範, 廣安 知之:【IT用語】Folksonomy ~タグで繋がるみんなの分類~:
<http://mikilab.doshisha.ac.jp/dia/research/report/2005/0911/001/report20050911001.html>:2005/10/24
- [2] taku-ku:MeCab Yet Another Part-of-Speech and Morphological Analyzer:
<http://mecab.sourceforge.net/>
- [3] Yahoo:Yahoo!デベロッパネットワーク:
<http://developer.yahoo.co.jp/>